interactions between such groups and the water in which they are, or otherwise would be, dissolved. The solvation shell (a highly ordered, and therefore thermodynamically disfavored, arrangement of water molecules around a non-polar group) around a single residue is reduced when another non-polar residue becomes positioned nearby during folding, releasing water in the solvation shell into the bulk solvent and thereby increasing the entropy of water solvent. It is estimated that approximately one-third of the ordered water molecules in an unfolded protein's solvation shell are lost into the bulk solvent upon formation of a secondary structure, and that about another one-third of original solvation water molecules are lost when a protein having a secondary structure folds into its tertiary structure.

Amino acid residues preferring hydrophobic environments tend to be "buried," *i.e.*, those found at least about 95% of the time within the interior of a folded protein, although positioning on the exterior surface of a globular protein can occur by placing the more polar components of the amino acid near the exterior surface. The clustering of two or more non-polar side chains on the exterior surface are generally associated with a biological function, *e.g.*, a substrate or ligand binding site. Polar amino acids are typically found on the exterior surface of globular proteins, where water stabilizes the residue's polarity. Positioning of an amino acid having a charged side chain in a globular protein's interior typically correlates with a structural or functional role for that residue with respect to biological function of the protein.

Another important protein folding parameter concerns hydrogen bond formation. A hydrogen bond (having bonding energies between about 1 to about 7 kcal/mol) is formed through the sharing of a hydrogen atom between two electronegative atoms, to one of which the hydrogen is covalently bonded (the hydrogen bond "donor"). Hydrogen bond strength depends primarily on the distance between the hydrogen bond donor and acceptor atoms, with high bond energies occurring when the donor and acceptor atoms are from about 2.7 Å to about 3.1 Å apart. Also contributing to hydrogen bond strength is bond geometry. Bonds

having high energies typically have the donor, hydrogen, and acceptors disposed in a colinear fashion. The dielectric constant of the medium surrounding the bond can also influence bond strength.

Electrostatic interactions (positive and negative) between charged amino acid residues also play a role in protein folding and substrate binding. The strength of these interactions varies directly with the charge on each ion and inversely with the solvent's dielectric constant and distance between the charges.

Other forces to consider in protein folding concern van der Waals forces, which involve both attractive and repulsive forces that depend on the distances between atoms. Attraction is believed to occur through induction of a complementary dipole in the electron density of adjacent atoms when electron orbitals approach at close distances. The repulsive component, also called steric hindrance, occurs at closer distances when neighboring atoms' electron orbitals begin to overlap. With regard to these forces, the most favorable interaction occurs at the van der Waals distance, which is the sum of the van der Waals radii for the two atoms. Van der Waals distances range from about 2.8 Å to about 4.1 Å. While individual van der Waals interactions usually have an energy less than 1 kcal/mol, the sum of these energies for even a protein of modest size is significant, and thus these interactions significantly impact protein folding and stability, and, ultimately, function.

Yet another interaction playing a role in protein folding and function concerns that which occurs when two or more aromatic rings approach each other such that the plane of the π electron orbitals of the aromatic rings overlap. Such interactions can have attractive, non-covalent forces of up to about 6 kcal/mol.

Other factors to consider in determining folding of proteins include the presence or absence of co-factors such as metals (*e.g.*, Zn^{2+}, Ca^{2+}, *etc.*), as well as other consideration known in the art.

Thermodynamic and kinetic considerations control the protein folding process. Without being tied to a particular theory, it is believed that folding begins

through short-range non-covalent interactions between several adjacent (as determined by primary structure) amino acid side chain groups and the polypeptide chain to which they are covalently linked. These interactions initiate folding of small regions of secondary structure, as certain R groups have a propensity to form α-helices, β structures, and sharp turns or bends in the protein backbone. Medium and long-range interactions between more distant regions of the protein then come into play as these distant regions become more proximate as the protein folds.

ALIGNMENT TECHNIQUES

Many sequence alignment methods are known in the art, such as BLAST (Altschul *et al.*, 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), and FASTA (Pearson & Lipman, 1988). Alignment methods such as these are typically employed to align amino acid sequences in order to determine the extent of amino acid sequence identity between an experimental, or "probe" or "target" amino acid sequence and one or more already stored sequences (the "template" amino acids sequence(s)).

Homology modeling can also be applied, particularly for amino acid sequences that are evolutionarily related, *i.e.*, they are homologous, such that their residue sequences can be aligned with some confidence. In one example of this method, the sequence of a protein whose structure has not been experimentally determined can be aligned to the sequence of a protein whose structure is known using one of the standard sequence alignment algorithms (Altschul, *et al.* (1990), *J. Mol. Biol., vol.* 215:403-410; Needleman and Wunsch (1970), *J. Mol. Biol., vol.* 48:443-453; Pearson and Lipman (1988), *Proc. Natl. Acad. Sci. USA, vol.* 85:2444-2448). Homology modeling algorithms, for example, Homology (Molecular Simulations, Inc.), build the sequence of the protein whose structure is not known onto the structure of the known protein to produce a "homology model".

An alternative approach to amino acid sequence alignment involves "threading" or "inverse folding" approaches. In such methods, one "threads" a